Research Paper

Title: Big Data Analysis of Petroleum Product Consumption in India

Prepared by: Kunal Sahu

1. ABSTRACT:

India's energy sector heavily depends on petroleum products that drive industrial, commercial, and transportation growth. This study applies **Big Data Analytics using Apache Spark (PySpark)** to analyze the monthly consumption of petroleum products in India.
The dataset, obtained from government sources, includes product-wise consumption (in thousand metric tonnes) across multiple years. Data preprocessing, aggregation, and visualization were conducted to identify annual trends, dominant fuel types, and consumption share by product category.
Results indicate that **High-Speed Diesel (HSD)** and **Motor Spirit (MS)** constitute the largest share of consumption, followed by **LPG**. A steady increase in demand was observed from 2020 to 2022, with a slight decline in 2023, likely due to rising EV usage and efficiency improvements. The project demonstrates how PySpark can efficiently handle and analyze largescale energy datasets for policy and forecasting purposes.

2. INTRODUCTION:

Petroleum products form the backbone of India's energy consumption. Tracking their usage trends is crucial for economic planning, import management, and sustainability goals.
Traditional analysis methods are inadequate for handling the massive size and complexity of government energy datasets. **Big Data Analytics** enables scalable data handling, faster insights, and better forecasting.
This paper presents a PySpark-based analysis of India's petroleum product consumption to uncover temporal trends and product-wise contributions.

3. OBJECTIVES:

1. To analyze yearly and product-wise petroleum consumption trends using PySpark.
2. To identify the most and least consumed petroleum products.
3. To study changes in total consumption over multiple years.
4. To visualize product dominance using Big Data visualization tools.

4. LITERATURE REVIEW:

Previous studies have shown that India's dependence on petroleum products remains high due to rapid industrialization and transportation growth.

- **Petroleum Planning & Analysis Cell (PPAC, 2023):** Reports an annual growth rate of 4–5% in overall petroleum demand.
- **IEA (2022):** Notes a gradual transition toward cleaner fuels but persistent dominance of diesel and petrol.
- **Energy Statistics India (2024):** Emphasizes the need for Big Data systems for real-time monitoring of petroleum distribution and consumption.

These studies highlight the importance of data-driven energy management — motivating the use of PySpark for analytical efficiency.

5. METHODOLOGY:

Data Source:
Government dataset — *Monthly Consumption of Petroleum Products in India (2020–2023)*

Column	Type	Description
Month	String	Name of month
Year	Integer	Year of record
PRODUCTS	String	Type of petroleum product (e.g., HSD, MS,LPG)
Quantity (000 Metric Tonnes)	Double	Consumption quantity

Tools & Technologies Used:

- **Big Data Framework:** Apache Spark (PySpark)
- **Programming Language:** Python
- **Visualization Tools:** Matplotlib, Seaborn
- **Environment:** Jupyter Notebook / Colab

Workflow:

1. **Data Ingestion:** Loading CSV dataset using PySpark.
2. **Data Cleaning:** Standardizing month names, handling null values, and converting quantities to numeric.
3. **Transformation:** Creating date fields and grouping by product and year.
4. **Analysis:** Yearly and product-wise aggregations using Spark SQL and DataFrame APIs.
5. **Visualization:** Generating line, bar, stacked area, and pie charts.
6. **Interpretation:** Identifying dominant fuels and annual consumption changes.

6. DATA ANALYSIS AND RESULTS:

1 Yearly Total Consumption: A continuous increase was observed from 2020 to 2022, showing post-COVID industrial recovery. However, 2023 showed a minor dip.

2 Product-Wise Contribution:

- **HSD (High-Speed Diesel)** → Highest consumption (~45–50% share)
- **MS (Motor Spirit)** → Second major contributor (~25%) • **LPG** → Significant household and industrial usage (~15%)

3 Visual Findings:

- **Line Chart:** Yearly total consumption trend shows steady growth with slight fluctuation.
- **Bar Chart:** HSD, MS, and LPG dominate total consumption.
- **Pie Chart:** Visualizes product share in total petroleum demand.
- **Area Chart:** Depicts top five products' yearly consumption trends.

7. INSIGHTS:

- Diesel and petrol continue to dominate India's fuel mix, highlighting transportation dependency.
- LPG consumption remains strong, reflecting its role in households and rural regions.
- The post-pandemic years show recovery in industrial activity and mobility.
- The slight decline in 2023 aligns with energy efficiency programs and electric vehicle adoption.

9. RECOMMENDATIONS:

- Promote **data-driven forecasting models** for energy demand prediction.
- Encourage **use of renewable alternatives** to reduce dependency on fossil fuels.
- Implement **real-time data collection systems** in petroleum distribution networks.
- Expand **Big Data infrastructure** for national energy analytics.

8. CONCLUSION:

The research demonstrates the potential of **Big Data Analytics using PySpark** for energy data processing and decision support.
Spark's distributed computing capabilities allowed efficient handling of thousands of records, while visual analytics provided actionable insights.
The results can help policymakers and energy planners anticipate demand and plan imports, subsidies, and sustainable energy transitions.

10. REFERENCES:

1. Petroleum Planning & Analysis Cell (PPAC, 2023). *Petroleum Statistics of India.*
2. International Energy Agency (IEA, 2022). *World Energy Outlook.*
3. Ministry of Petroleum and Natural Gas (2023). *Monthly Consumption Reports.*
4. Apache Software Foundation. *PySpark Documentation.*